

# Segmentation-driven 6D Object Pose Estimation

Yinlin Hu, Joachim Hugonot, Pascal Fua, Mathieu Salzmann

CVLab, EPFL, Switzerland

{yinlin.hu, joachim.hugonot, pascal.fua, mathieu.salzmann}@epfl.ch

## Abstract

*The most recent trend in estimating the 6D pose of rigid objects has been to train deep networks to either directly regress the pose from the image or to predict the 2D locations of 3D keypoints, from which the pose can be obtained using a PnP algorithm. In both cases, the object is treated as a global entity, and a single pose estimate is computed. As a consequence, the resulting techniques can be vulnerable to large occlusions.*

*In this paper, we introduce a segmentation-driven 6D pose estimation framework where each visible part of the objects contributes a local pose prediction in the form of 2D keypoint locations. We then use a predicted measure of confidence to combine these pose candidates into a robust set of 3D-to-2D correspondences, from which a reliable pose estimate can be obtained. We outperform the state-of-the-art on the challenging Occluded-LINEMOD and YCB-Video datasets, which is evidence that our approach deals well with multiple poorly-textured objects occluding each other. Furthermore, it relies on a simple enough architecture to achieve real-time performance.*

## 1. Introduction

Image-based 6D object pose estimation is crucial in many real-world applications, such as augmented reality or robot manipulation. Traditionally, it has been handled by establishing correspondences between the object's known 3D model and 2D pixel locations, followed by using the Perspective-n-Point (PnP) algorithm to compute the 6 pose parameters [19, 38, 43]. While very robust when the object is well textured, this approach can fail when it is featureless or when the scene is cluttered with multiple objects occluding each other.

Recent work has therefore focused on overcoming these difficulties, typically using deep networks to either regress directly from image to 6D pose [17, 45] or to detect keypoints associated to the object [35, 39], which can then be used to perform PnP. In both cases, however, the object is

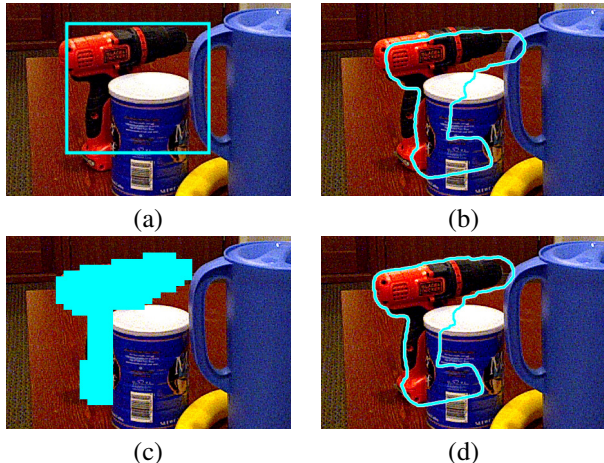


Figure 1: **Global pose estimation vs our segmentation-driven approach.** (a) The drill's bounding box overlaps another occluding it. (b) As a result, the globally-estimated pose [45] is wrong. (c) In our approach, only image patches labeled as corresponding to the drill contribute to the pose estimate. (d) It is now correct.

still treated as a global entity, which makes the algorithm vulnerable to large occlusions. Fig. 1 depicts a such a case: The bounding box of an occluded drill overlaps other objects that provide irrelevant information to the pose estimator and thereby degrade its performance. Because this happens often, many of these recent methods require an additional post-processing step to refine the pose [23].

In this paper, we show that more robust pose estimates can be obtained by combining multiple local predictions instead of a single global one. To this end, we introduce a segmentation-driven 6D pose estimation network in which each visible object patch contributes a pose estimate for the object it belongs to in the form of the predicted 2D projections of predefined 3D keypoints. Using confidence values also predicted by our network, we then combine the most reliable 2D projections for each 3D keypoint, which yields a robust set of 3D-to-2D correspondences. We then use a RANSAC-based PnP strategy to infer a single reliable pose per object.

Reasoning in terms of local patches not only makes our

approach robust to occlusions, but also yields a rough segmentation of each object in the scene. In other words, unlike other methods that divorce object detection from pose estimation [35, 17, 45], we perform both jointly while still relying on a simple enough architecture for real-time performance.

In short, our contribution is a simple but effective segmentation-driven network that produces accurate 6D object pose estimates without the need for post-processing, even when there are multiple poorly-textured objects occluding each other. It combines segmentation and ensemble learning in an effective and efficient architecture. We will show that it outperforms the state-of-the-art methods on standard benchmarks, such as the OccludedLINEMOD and YCB-Video datasets.

## 2. Related Work

In this paper, we focus on 6D object pose estimation from RGB images, without access to a depth map, unlike in RGBD-based methods [12, 2, 3, 28]. The classical approach to performing this task involves extracting local features from the input image, matching them with those of the model, and then running a PnP algorithm on the resulting 3D-to-2D correspondences. Over the years, much effort has been invested in designing local feature descriptors that are invariant to various transformations [27, 40, 41, 42, 34, 32], so that they can be matched more robustly [29, 30, 14]. In parallel, increasingly effective PnP methods have been developed to handle noise and mismatches [20, 46, 22, 8]. As a consequence, when dealing with well-textured objects, feature-based pose estimation is now fast and robust, even in the presence of mild occlusions. However, it typically struggles with heavily-occluded and poorly-textured objects.

In the past, textureless objects have often been handled by template-matching [11, 12]. Image edges then become the dominant information source [21, 26], and researchers have developed strategies based on different distances, such as the Hausdorff [15] and the Chamfer [25, 13] ones, to match the 3D model against the input image. While effective for poorly-textured objects, these techniques often fail in the presence of mild occlusions and cluttered background.

As in many computer vision areas, the modern take on 6D object pose estimation involves deep neural networks. Two main trends have emerged: Either regressing from the image directly to the 6D pose [17, 45] or predicting 2D keypoint locations in the image [35, 39], from which the pose can be obtained via PnP. Both approaches treat the object as a global entity and produce a single pose estimate. This makes them vulnerable to occlusions because, when considering object bounding boxes as they all do, signal coming from other objects or from the background will contaminate the prediction. While, in [35, 45], this is addressed by seg-

menting the object of interest, the resulting algorithms still provide a single, global pose estimate, that can be unreliable, as illustrated in Fig. 1 and demonstrated in the results section. As a consequence, these methods typically invoke an additional pose refinement step [23].

To the best of our knowledge, the work of [16] and [31] constitute the only recent attempts at going beyond a global prediction. While the method in [16] also relies on segmentation via a state-of-the-art semantic segmentation network, its use of regression to 3D object coordinates, which reside in a very large space, yields disappointing performance. By contrast, the technique in [31] predicts multiple keypoint location heatmaps from local patches and assembles them to form an input to a PnP algorithm. The employed patches, however, remain relatively large, thus still potentially containing irrelevant information. Furthermore, at runtime, this approach relies on a computationally-expensive sliding-window strategy that is ill-adapted to real-time processing. Here, we propose to achieve robustness by combining multiple local pose predictions in an ensemble manner and in real time, without post-processing. In the results section, we will show that this outperforms the state-of-the-art approaches [17, 45, 35, 39, 16, 31].

Note that human pose estimation [10, 44, 33] is also related to global 6D object pose prediction techniques. By targeting non-rigid objects, however, these methods require the more global information extracted from larger receptive fields and are inevitably more sensitive to occlusions. By contrast, dealing with rigid objects allows us to rely on local predictions that can be robustly combined, and local visible object parts can provide reliable predictions for all keypoints. We show that assembling these local predictions yields robust pose estimates, even when observing multiple objects that occlude each other.

## 3. Approach

Given an input RGB image, our goal is to simultaneously detect objects and estimate their 6D pose, in terms of 3 rotations and 3 translations. We assume the objects to be rigid and their 3D model to be available. As in [35, 39], we design a CNN architecture to regress the 2D projections of some predefined 3D points, such as the 8 corners of the objects' bounding boxes. However, unlike these methods whose predictions are global for each object and therefore affected by occlusions, we make individual image patches predict both to which object they belong and where the 2D projections are. We then combine the predictions of all patches assigned to the same object for robust PnP-based pose estimation.

Fig. 2 depicts the corresponding workflow. In the remainder of this section, we first introduce our two-stream network architecture. We then describe each stream individually and finally our inference strategy.

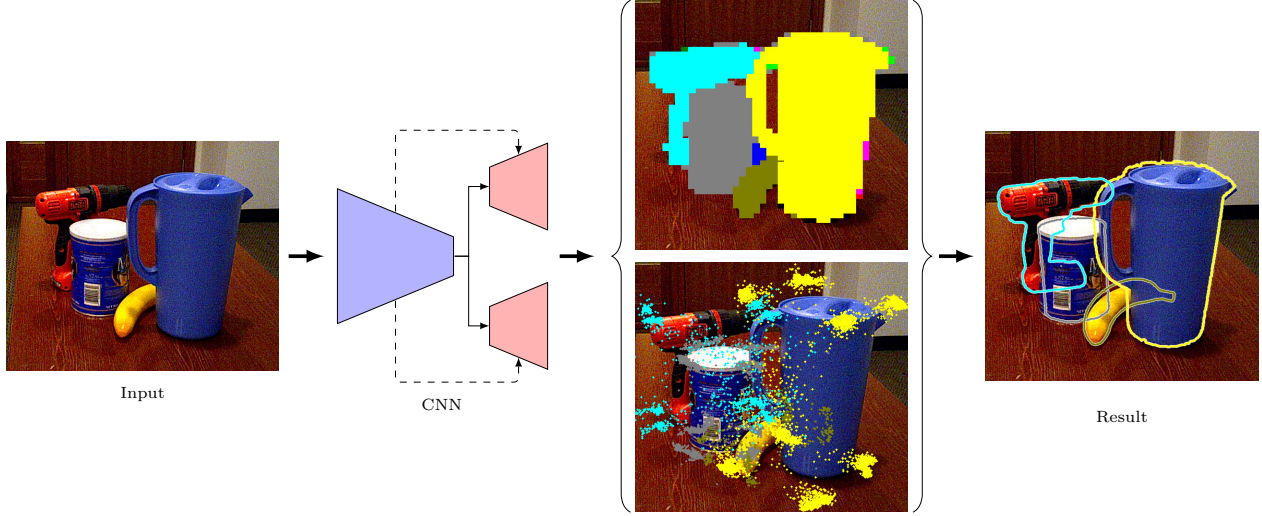


Figure 2: **Overall workflow of our method.** Our architecture has two streams: One for object segmentation and the other to regress 2D keypoint locations. These two streams share a common encoder, but the decoders are separate. Each one produces a tensor of a spatial resolution that defines an  $S \times S$  grid over the image. The segmentation stream predicts the label of the object observed at each grid location. The regression stream predicts the 2D keypoint locations for that object.

### 3.1. Network Architecture

In essence, we aim to jointly perform segmentation by assigning image patches to objects and 2D coordinate regression of keypoints belonging to these objects, as shown in Fig. 3. To this end, we design the two-stream architecture depicted by Fig. 2, with one stream for each task. It has an encoder-decoder structure, with a common encoder for both streams and two separate decoders.

For the encoder, we use the Darknet-53 architecture of YOLOv3 [37] that has proven highly effective and efficient for objection detection. For the decoders, we designed networks that output 3D tensors of spatial resolution  $S \times S$  and feature dimensions  $D_{seg}$  and  $D_{reg}$ , respectively. This amounts to superposing an  $S \times S$  grid on the image and computing a feature vector of dimension  $D_{seg}$  or  $D_{reg}$  per grid element. The spatial resolution of that grid controls the size of the image patches that vote for the object label and specific keypoint projections. A high resolution yields fine segmentation masks and many votes. However, it comes at a higher computational cost, which may be unnecessary for our purposes. Therefore, instead of matching the 5 down-sampling layers of the Darknet-53 encoder with 5 upsampling layers, we only use 2 such layers, with a standard stride of 2. The same architecture, albeit with a different output feature size, is used for both decoder streams.

To train our model end-to-end, we define a loss function

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{reg}, \quad (1)$$

which combines a segmentation and a regression term that we use to score the output of each stream. We now turn to their individual descriptions.

### 3.2. Segmentation Stream

The role of the segmentation stream is to assign a label to each cell of the virtual  $S \times S$  grid superposed on the image, as shown in Fig. 3(a). More precisely, given  $K$  object classes, this translates into outputting a vector of dimension  $D_{seg} = K + 1$  at each spatial location, with an additional dimension to account for the background.

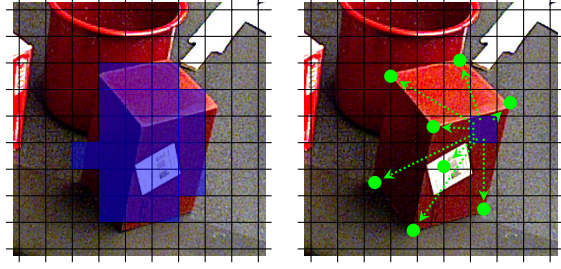
During training, we have access to both the 3D object models and their ground-truth pose. We can therefore generate the ground-truth semantic labels by projecting the 3D models in the images while taking into account the depth of each object to handle occlusions. In practice, the images typically contain many more background regions than object ones. Therefore, we take the loss  $\mathcal{L}_{seg}$  of Eq. 1 to be the Focal Loss of [24], a dynamically weighted version of the cross-entropy. Furthermore, we rely on the median frequency balancing technique of [6, 1] to weigh the different samples. We do this according to the pixel-wise class frequencies rather than the global class frequencies to account for the fact that objects have different sizes.

### 3.3. Regression Stream

The purpose of the regression stream is to predict the 2D projections of predefined 3D keypoints associated to the 3D object models. Following standard practice [35, 9, 36], we typically take these keypoints to be the 8 corners of the model bounding boxes.

Recall that the output of the regression stream is a 3D tensor of size  $S \times S \times D_{reg}$ . Let  $N$  be the number of 3D keypoints per object whose projection we want to predict. When using bounding box corners,  $N = 8$ . We take  $D_{reg}$





(a) Object segmentation. (b) Keypoint 2D locations.

Figure 3: **Outputs of our two-stream network.** (a) The segmentation stream assigns a label to each cell of the virtual grid superposed on the image. (b) In the regression stream, each grid cell predicts the 2D keypoint locations of the object it belongs to. Here, we take the 8 bounding box corners to be our keypoints.

to be  $3N$  to represent at each spatial location the  $N$  pairs of 2D projection values along with a confidence value for each.

In practice, we do not predict directly the keypoints' 2D coordinates. Instead, for each one, we predict an offset vector with respect to the center of the corresponding grid cell, as illustrated by Fig. 3(b). That is, let  $\mathbf{c}$  be the 2D location of a grid cell center. For the  $i^{th}$  keypoint, we seek to predict an offset  $\mathbf{h}_i(\mathbf{c})$ , such that the resulting location  $\mathbf{c} + \mathbf{h}_i(\mathbf{c})$  is close to the ground-truth 2D location  $\mathbf{g}_i$ . During training, this is expressed by the residual

$$\Delta_i(\mathbf{c}) = \mathbf{c} + \mathbf{h}_i(\mathbf{c}) - \mathbf{g}_i, \quad (2)$$

and by defining the loss function

$$\mathcal{L}_{pos} = \sum_{\mathbf{c} \in M} \sum_{i=1}^N \|\Delta_i(\mathbf{c})\|_1, \quad (3)$$

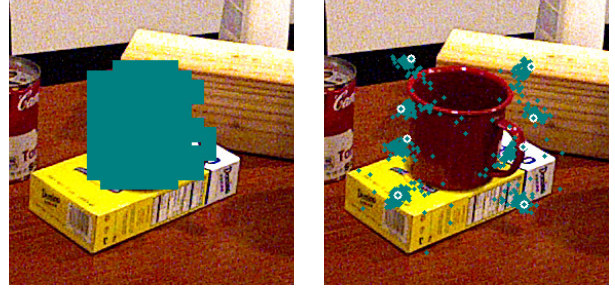
where  $M$  is the foreground segmentation mask, and  $\|\cdot\|_1$  denotes the  $L^1$  loss function, which is less sensitive to outliers than the  $L^2$  loss. Only accounting for the keypoints that fall within the segmentation mask  $M$  focuses the computation on image regions that truly belong to objects.

As mentioned above, the regression stream also outputs a confidence value  $s_i(\mathbf{c})$  for each predicted keypoint, which is obtained via a sigmoid function on the network output. These confidence values should reflect the proximity of the predicted 2D projections to the ground truth. To encourage this, we define a second loss term

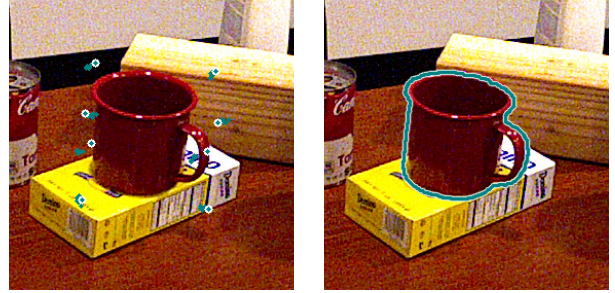
$$\mathcal{L}_{conf} = \sum_{\mathbf{c} \in M} \sum_{i=1}^N \|s_i(\mathbf{c}) - \exp(-\tau \|\Delta_i(\mathbf{c})\|_2)\|_1, \quad (4)$$

where  $\tau$  is a modulating factor. We then take the regression loss term of Eq. 1 to be

$$\mathcal{L}_{reg} = \beta \mathcal{L}_{pos} + \gamma \mathcal{L}_{conf}, \quad (5)$$



(a) (b)



(c) (d)

Figure 4: **Combining pose candidates.** (a) Grid cells predicted to belong to the cup are overlaid on the image. (b) Each one predicts 2D locations for the corresponding keypoints, shown as green dots. (c) For each 3D keypoint, the  $n = 10$  2D locations about which the network is most confident are selected. (d) Running a RANSAC-based PnP on these yields an accurate pose estimate, as evidenced by the correctly drawn outline.

where  $\beta$  and  $\gamma$  modulate the influence of the two terms.

Note that because the two terms in Eq. 5 focus on the regions that are within the segmentation mask  $M$ , their gradients are also backpropagated to these regions only. As in the segmentation stream, to account for pixel-wise class imbalance, we weigh the regression loss term for different objects according to the pixel-wise class frequencies in the training set.

### 3.4. Inference Strategy

At test time, given a query image, our network returns, for each foreground cell in the  $S \times S$  grid of Section 3.1, an object class and a set of predicted 2D locations for the projection of the  $N$  3D keypoints. As we perform class-based segmentation instead of instance-based segmentation, there might be ambiguities if two objects of the same class are present in the scene. To avoid that, we leverage the fact that the predicted 2D keypoint locations tend to cluster according to the objects they correspond and use a simple pixel distance threshold to identify such clusters.

For each cluster, that is, for each object, we then exploit the confidence scores predicted by the network to establish 2D-to-3D correspondences between the image and the object's 3D model. The simplest way of doing so would

	ADD-0.1d			REP-5px					
	PoseCNN	Heatmaps	Ours	PoseCNN	BB8	Tekin	iPose	Heatmaps	Ours
Ape	9.6	<b>16.5</b>	12.1	34.6	28.5	7.0	24.2	<b>64.7</b>	59.1
Can	<b>45.2</b>	42.5	39.9	15.1	1.2	11.2	30.2	53.0	<b>59.8</b>
Cat	0.9	2.8	<b>8.2</b>	10.4	9.6	3.6	12.3	<b>47.9</b>	46.9
Driller	41.4	<b>47.1</b>	45.2	7.4	0	1.4	-	35.1	<b>59.0</b>
Duck	<b>19.6</b>	11.0	17.2	31.8	6.8	5.1	12.1	36.1	<b>42.6</b>
Eggbox*	22.0	<b>24.7</b>	22.1	1.9	-	-	-	10.3	<b>11.9</b>
Glue*	38.5	<b>39.5</b>	35.8	13.8	4.7	6.5	25.9	<b>44.9</b>	16.5
Holepun.	22.1	21.9	<b>36.0</b>	23.1	2.4	8.3	20.6	52.9	<b>63.6</b>
Average	24.9	25.8	<b>27.0</b>	17.2	7.6	6.2	20.8	43.1	<b>44.9</b>

Table 1: **Comparison with the state of the art on Occluded-LINEMOD.** We compare our results with those of PoseCNN [45], BB8 [35], Tekin [39], iPose [16], and Heatmaps [31]. The results missing from the original papers are denoted as “-”.

	PoseCNN	BB8	Tekin	Heatmaps	iPose	Ours
FPS	4	3	50	4	-	22

Table 2: **Runtime comparisons on Occluded-LINEMOD.** All methods run on a modern Nvidia GPU.

	NF	HC	B-2	B-10	Oracle
Ape	37.8	58.2	58.2	59.1	<b>84.0</b>
Can	53.4	58.7	58.5	59.8	<b>89.0</b>
Cat	42.6	46.1	47.4	46.9	<b>60.6</b>
Driller	52.5	56.8	59.4	59.0	<b>90.3</b>
Duck	40.4	42.8	42.4	42.6	<b>55.6</b>
Eggbox*	<b>12.8</b>	11.2	12.1	11.9	10.9
Glue*	14.7	15.8	15.1	16.5	<b>41.0</b>
Holepun.	58.4	62.2	63.1	63.6	<b>89.3</b>
Average	39.1	44.0	44.5	44.9	<b>65.1</b>
FPS	26	26	25	22	-

Table 3: **Accuracy (REP-5px) of different fusion strategies on Occluded-LINEMOD.** We compare a No-Fusion (NF) scheme with one that relies on the Highest-Confidence predictions, and with strategies relying on performing RANSAC on the  $n$  most confident predictions (B- $n$ ). Oracle consists of choosing the best 2D location using the ground-truth one, and is reported to indicate the potential for improvement of our approach. In the bottom row, we also report the average runtime of these different strategies.

be to use RANSAC on all the predictions. This, however, would significantly slow down our approach. Instead, we rely on the  $n$  most confident 2D predictions for each 3D keypoint. In practice, we found  $n = 10$  to yield a good balance between speed and accuracy. Given these filtered 2D-to-3D correspondences, we obtain the 6D pose of each object using the RANSAC-based version of the EPnP algorithm of [20]. Fig. 4 illustrates this procedure.

## 4. Experiments

We now evaluate our segmentation-driven multi-object 6D pose estimation method on the challenging Occluded-LINEMOD [18] and YCB-Video [45] datasets, which, unlike LINEMOD [12], contain 6D pose annotations for each object appearing in all images.

**Metrics.** We report the commonly-used 2D reprojection (REP) error [3]. It encodes the average distance between the 2D reprojection of the 3D model points obtained using the predicted pose and those obtained with the ground-truth one. Furthermore, we also report the pose error in 3D space [12], which corresponds to the average distance between the 3D points transformed using the predicted pose and those obtained with the ground-truth one. As in [23, 45], we will refer to it as ADD. Since many objects in the datasets are symmetric, we use the symmetric version of these two metrics and report their REP-5px and ADD-0.1d values. They assume the predicted pose to be correct if the REP is below a 5 pixel threshold and the ADD below 10% of the model diameter, respectively. Below, we denote the objects that are considered to be symmetric by a \* superscript.

**Implementation Details.** As in [37], we scale the input image to a  $608 \times 608$  resolution for both training and testing. Furthermore, when regressing the 2D reprojections, we normalize the horizontal and vertical positions to the range  $[0, 10]$ . We use the same normalization procedure when estimating the confidences.

We train the network for 300 epochs on Occluded-LINEMOD and 30 epochs on YCB-Video. In both cases, the initial learning rate is set to  $1e-3$ , and is divided by 10 after 50%, 75%, and 90% of the total number of epochs. We use SGD as our optimizer with a momentum of 0.9 and a weight decay of  $5e-4$ . Each training batch contains 8 images, and we have employed the usual data augmentation techniques, such as random luminance, Gaus-

	ADD-0.1d				REP-5px					
	Mask R-CNN	CPM	[45]	Ours	Mask R-CNN	CPM	[45]	[35]	[39]	Ours
Average	11.8	12.7	24.9	<b>27.0</b>	22.4	22.9	17.2	7.6	6.2	<b>44.9</b>

Table 4: **Comparison with human pose estimation methods on Occluded-LINEMOD.** We modified two state-of-the-art human pose estimation methods, Mask R-CNN [10] and CPM [44], to output bounding box corner locations. While both Mask R-CNN and CPM perform slightly better than other global-inference methods, our local approach yields much more accurate predictions.

sian noise, translation and scaling. We have also used the random erasing technique of [47] for better occlusion handling. Our source code is publicly available at <https://github.com/cvlab-epfl/segmentation-driven-pose>.

#### 4.1. Evaluation on Occluded-LINEMOD

The Occluded-LINEMOD dataset [18] was compiled by annotating the pose of all the objects in a subset of the raw LINEMOD dataset [12]. This subset depicts 8 different objects in 1214 images. Although depth information is also provided, we only exploit the RGB images. The Occluded-LINEMOD images, as the LINEMOD ones, depict a central object surrounded by non-central ones. The standard protocol consists of only evaluating on the *non-central* objects.

To create training data for our model, we follow the same procedure as in [23, 39]. We use the mask inferred from the ground-truth pose to segment the *central* object in each image, since, as mentioned above, it will not be used for evaluation. We then generate synthetic images by inpainting between 3 and 8 objects on random PASCAL VOC images [7]. These objects are placed at random locations, orientations, and scales. This procedure still enables us to recover the occlusion state of each object and generate the corresponding segmentation mask. By using the central objects from any of the raw LINEMOD images, provided that it is one of the 8 objects used in Occluded-LINEMOD, we generated 20k training samples.

##### 4.1.1 Comparing against the State of the Art

We compare our method with the state-of-the-art ones of [45] (PoseCNN), [35] (BB8), and [39] (Tekin), which all produce a single global pose estimate. Furthermore, we also report the results of the recent work of [16] (iPose), and [31] (Heatmaps), which combines the predictions of multiple, relatively large patches, but relies on an expensive sliding-window strategy. Note that [31] also provides results obtained with the Feature Mapping technique [36]. However, most methods, including ours, do not use this technique, and for a fair comparison, we therefore report the results of all methods, including that of [31], without it.

We report our results in Table 1 and provide the runtimes of the methods in Table 2. Our method outperforms the global inference ones [45, 35, 39] by a large margin. It also outperforms Heatmaps, albeit by a smaller one. Further-

more, thanks to our simple architecture and one-shot inference strategy, our method runs more than 5 times faster than Heatmaps. Our approach takes 30ms per-image for segmentation and 2D reprojection estimation, and 3-4ms per object for fusion. With 5 objects per image on average, this yields a runtime of about 50ms. Fig. 5 depicts some of our results. Note their accuracy even in the presence of large occlusions.

##### 4.1.2 Comparison of Different Fusion Strategies

As shown in Fig. 4, not all local predictions of the 2D keypoint locations are accurate. Therefore, the fusion strategy based on the predicted confidence values that we described in Section 3.4 is important to select the right ones. Here, we evaluate its impact on the final pose estimate. To this end, we report the results obtained by taking the 2D location with highest confidence (HC) for each 3D keypoint and those obtained with different values  $n$  in our  $n$  most-confident selection strategy. We refer to this as B- $n$  for a particular value  $n$ . Note that we then use RANSAC on the selected 2D-to-3D correspondences.

In Table 3, we compare the results of these different strategies with a fusion-free method that always uses the 2D reprojections predicted by the center grid, which we refer to as No-Fusion (NF). These results evidence that all fusion schemes outperform the No-Fusion one. We also report the Oracle results obtained by selecting the best predicted 2D location for each 3D keypoint using the ground truth 2D reprojections. This indicates that our approach could further benefit from improving the confidence predictions or designing a better fusion scheme.

##### 4.1.3 Comparison with Human Pose Methods

Our method enables us to infer keypoints' locations of rigid objects from local visible object regions and does not require the more global information extracted from larger receptive fields that are more sensitive to occlusions. To further back up this claim, we compare our approach to two state-of-the-art human pose estimation methods, Mask R-CNN [10] and Convolutional Pose Machines (CPM) [44], which target non-rigid objects, *i.e.* human bodies. By contrast, dealing with rigid objects allows us to rely on local predictions that can be robustly combined. Specifically, we modified the publicly available code of Mask R-CNN



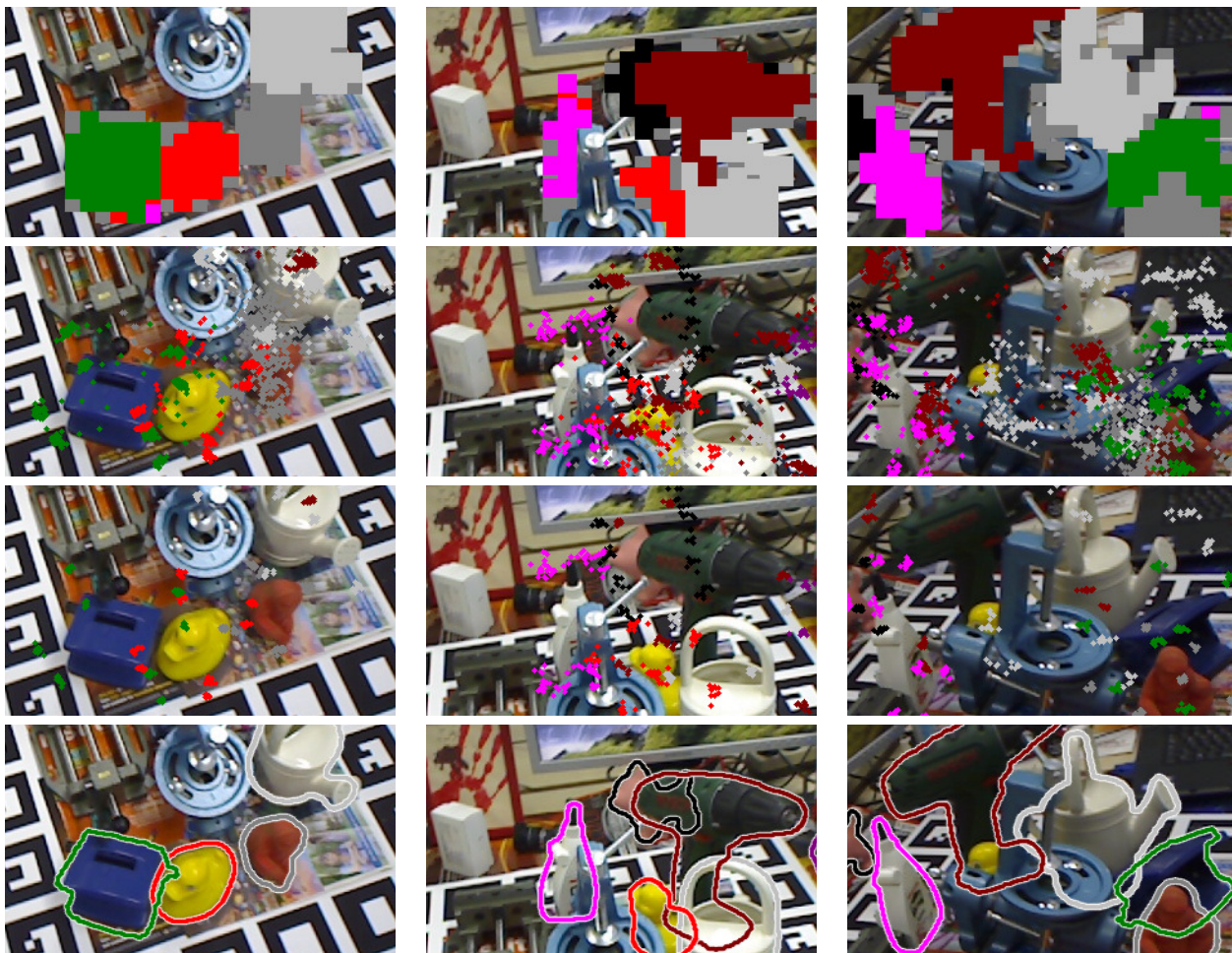


Figure 5: **Occluded-LINEMOD results.** In each column, we show, from top to bottom: the foreground segmentation mask, all 2D reprojection candidates, the selected 2D reprojections, and the final pose results. Our method generates accurate pose estimates, even in the presence of large occlusions. Furthermore, it can process multiple objects in real time.



Figure 6: **Comparison to PoseCNN [45] on YCB-Video.** (Top) PoseCNN and (Bottom) Our method. This demonstrates the benefits of reasoning about local object parts instead of globally, particularly in the presence of large occlusions.

	ADD-0.1d			REP-5px		
	[45]	[31]	Ours	[45]	[31]	Ours
master_chef_can	3.6	32.9	<b>33.0</b>	0.1	9.9	<b>21.0</b>
cracker_box	25.1	<b>62.6</b>	44.6	0.1	<b>24.5</b>	12.0
sugar_box	40.3	44.5	<b>75.6</b>	7.1	47.0	<b>56.3</b>
tomato_soup_can	25.5	31.1	<b>40.8</b>	5.2	41.5	<b>46.2</b>
mustard_bottle	61.9	42.0	<b>70.6</b>	6.4	42.3	<b>70.3</b>
tuna_fish_can	11.4	6.8	<b>18.1</b>	3.0	7.1	<b>39.3</b>
pudding_box	14.5	<b>58.4</b>	12.2	5.1	<b>43.9</b>	17.3
gelatin_box	12.1	42.5	<b>59.4</b>	15.8	62.1	<b>83.6</b>
potted_meat_can	18.9	<b>37.6</b>	33.3	23.1	38.5	<b>60.7</b>
banana	<b>30.3</b>	16.8	16.6	0.3	8.2	<b>22.4</b>
pitcher_base	15.6	57.2	<b>90.0</b>	0	15.9	<b>33.5</b>
bleach_cleanser	21.2	65.3	<b>70.9</b>	1.2	12.1	<b>43.3</b>
bowl*	12.1	25.6	<b>30.5</b>	4.4	<b>16.0</b>	13.3
mug	5.2	11.6	<b>40.7</b>	0.8	20.3	<b>38.1</b>
power_drill	29.9	46.1	<b>63.5</b>	3.3	40.9	<b>43.3</b>
wood_block*	10.7	<b>34.3</b>	27.7	0	<b>2.5</b>	<b>2.5</b>
scissors	2.2	0	<b>17.1</b>	0	0	<b>8.8</b>
large_marker	3.4	3.2	<b>4.8</b>	1.4	0	<b>13.6</b>
large_clamp*	<b>28.5</b>	10.8	25.6	0.3	0	<b>7.6</b>
extra_large_clamp*	19.6	<b>29.6</b>	8.8	<b>0.6</b>	0	<b>0.6</b>
foam_brick*	<b>54.5</b>	51.7	34.7	0	<b>52.4</b>	13.5
Average	21.3	33.6	<b>39.0</b>	3.7	23.1	<b>30.8</b>

Table 5: **Comparison with the state of the art on YCB-Video.** We compare our results with those of PoseCNN [45] and Heatmaps [31].

and CPM to output 8 bounding box 2D corners instead of human keypoints and trained these methods on Occluded-LINEMOD. As shown in Table 4, while both Mask R-CNN and CPM perform slightly better than other *global-inference* methods, our *local* approach yields much more accurate predictions.

## 4.2. Evaluation on YCB-Video

We also evaluate our method on the recent and more challenging YCB-Video dataset [45]. It comprises 21 objects taken from the YCB dataset [5, 4], which are of diverse sizes and with different degrees of texture. This dataset contains about 130K real images from 92 video sequences, with an additional 80K synthetically rendered images that only contain foreground objects. It provides the pose annotations of all the objects, as well as the corresponding segmentation masks. The test images depict a great diversity in illumination, noise, and occlusions, which makes this dataset extremely challenging. As before, while depth information is available, we only use the color images. Here, we generate complete synthetic images from the 80K synthetic foreground ones by using the same random background procedure as in Section 4.1. As before, we report results without feature mapping, because neither PoseCNN nor our approach use them.

### 4.2.1 Comparing against the State of the Art

Fewer methods have reported results on this newer dataset. In Table 5, we contrast our method with the two baselines that have. Our method clearly outperforms both PoseCNN [45] and Heatmaps [31]. Furthermore, recall that our approach runs more than 5 times faster than either of them.

In Fig. 6, we compare qualitative results of PoseCNN and ours. While our pose estimates are not as accurate on this dataset as on Occluded-LINEMOD, they are still much better than those of PoseCNN. Again, this demonstrates the benefits of reasoning about local object parts instead of globally, particularly in the presence of large occlusions.

## 4.3. Discussion

Although our method performs well in most cases, it still can handle neither the most extreme occlusions nor tiny objects. In such cases, the grid we rely on becomes too rough a representation. This, however, could be addressed by using a finer grid, or, to limit the computational burden, a grid that is adaptively subdivided to better handle each image region. Furthermore, as shown in Table 3, we do not yet match the performance of an oracle that chooses the best predicted 2D location for each 3D keypoint. This suggests that there is room to improve the quality of the predicted confidence score, as well as the fusion procedure itself. This will be the topic of our future research.

## 5. Conclusion

We have introduced a segmentation-driven approach to 6D object pose estimation, which jointly detects multiple objects and estimates their pose. By combining multiple local pose estimates in a robust fashion, our approach produces accurate results without the need for a refinement step, even in the presence of large occlusions. Our experiments on two challenging datasets have shown that our approach outperforms the state of the art, and, as opposed to the best competitors, predicts the pose of multiple objects in real time. In the future, we will investigate the use of other backbone architectures for the encoder and devise a better fusion strategy to select the best predictions before performing PnP. We will also seek to incorporate the PnP step of our approach into the network, so as to have a complete, end-to-end learning framework.

**Acknowledgments** This work was supported in part by the Swiss Innovation Agency Innosuisse. We would like to thank Markus Oberweger and Yi Li for clarifying details about their papers, and Zheng Dang for helpful discussions.



## References

- [1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv Preprint*, 2015. 3
- [2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D Object Pose Estimation Using 3D Object Coordinates. In *European Conference on Computer Vision*, 2014. 2
- [3] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5
- [4] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-CMU-Berkeley dataset for robotic manipulation research. In *International Journal of Robotics Research*, 2017. 8
- [5] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron Dollar. The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research. In *International Conference on Advanced Robotics*, 2015. 8
- [6] David Eigen and Rob Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. In *International Conference on Computer Vision*, 2015. 3
- [7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6
- [8] Luis Ferraz, Xavier Binefa, and Francesc Moreno-Noguer. Very Fast Solution to the PnP Problem with Algebraic Outlier Rejection. In *Conference on Computer Vision and Pattern Recognition*, pages 501–508, 2014. 2
- [9] Alexander Grabner, Peter M. Roth, and Vincent Lepetit. 3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *International Conference on Computer Vision*, 2017. 2, 6
- [11] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter F. Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient Response Maps for Real-Time Detection of Textureless Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, May 2012. 2
- [12] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary R. Bradski, Kurt Konolige, and Nassir Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *Asian Conference on Computer Vision*, 2012. 2, 5, 6
- [13] Edward Hsiao, Sudipta N. Sinha, Krishnan Ramnath, Simon Baker, C. Lawrence Zitnick, and Richard Szeliski. Car Make and Model Recognition Using 3D Curve Alignment. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 2
- [14] Yinlin Hu, Rui Song, and Yunsong Li. Efficient Coarse-to-Fine Patch Match for Large Displacement Optical Flow. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2
- [15] Daniel P. Huttenlocher, Gregory A. Klanderman, and William Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 850–863, 1993. 2
- [16] Omid Hosseini Jafari, Siva Karthik Mustikovela, Karl Pertsch, Eric Brachmann, and Carsten Rother. iPose: Instance-Aware 6D Pose Estimation of Partly Occluded Objects. In *Asian Conference on Computer Vision*, 2018. 2, 5, 6
- [17] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: Making Rgb-Based 3D Detection and 6D Pose Estimation Great Again. In *International Conference on Computer Vision*, 2017. 1, 2
- [18] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning Analysis-By-Synthesis for 6D Pose Estimation in RGB-D Images. In *International Conference on Computer Vision*, 2015. 5, 6
- [19] Vincent Lepetit and Pascal Fua. *Monocular Model-Based 3D Tracking of Rigid Objects: A Survey*. Now Publishers, September 2005. 1
- [20] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision*, 2009. 2, 5
- [21] Dengwang Li, Haiguang Wang, Yong Yin, and Xiuying Wang. Deformable Registration Using Edge-preserving Scale Space for Adaptive Image-guided Radiation Therapy. In *Journal of Applied Clinical Medical Physics*, 2011. 2
- [22] Shiqi Li, Chi Xu, and Ming Xie. A Robust O(n) Solution to the Perspective-N-Point Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1444–1450, 2012. 2
- [23] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Poseestimation. In *European Conference on Computer Vision*, 2018. 1, 2, 5, 6
- [24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *International Conference on Computer Vision*, 2017. 3
- [25] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, and Rama Chellappa. Fast Directional Chamfer Matching. In *Conference on Computer Vision and Pattern Recognition*, 2010. 2
- [26] David G. Lowe. Fitting Parameterized Three-Dimensional Models to Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, June 1991. 2
- [27] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 20(2):91–110, Nov 2004. 2
- [28] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global Hypothesis Generation for 6D Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2017. 2

- [29] Marius Muja and David G. Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In *International Conference on Computer Vision*, 2009. 2
- [30] Marius Muja and David G. Lowe. Scalable Nearest Neighbor Algorithms for High Dimensional Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 2
- [31] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. In *European Conference on Computer Vision*, 2018. 2, 5, 6, 8
- [32] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning Local Features from Images. In *Advances in Neural Information Processing Systems*, 2018. 2
- [33] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In *European Conference on Computer Vision*, 2018. 2
- [34] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, and Kostas Daniilidis. 6-DoF Object Pose from Semantic Keypoints. In *International Conference on Robotics and Automation*, 2017. 2
- [35] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *International Conference on Computer Vision*, 2017. 1, 2, 3, 5, 6
- [36] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images. In *Conference on Computer Vision and Pattern Recognition*, 2018. 3, 6
- [37] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. In *arXiv Preprint*, 2018. 3, 5
- [38] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints. *International Journal of Computer Vision*, 66(3), 2006. 1
- [39] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-Time Seamless Single Shot 6D Object Pose Prediction. In *Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6
- [40] Engin Tola, Vincent Lepetit, and Pascal Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010. 2
- [41] Tomasz Trzcinski, C. Mario Christoudias, Vincent Lepetit, and Pascal Fua. Learning Image Descriptors with the Boosting-Trick. In *Advances in Neural Information Processing Systems*, December 2012. 2
- [42] Shubham Tulsiani and Jitendra Malik. Viewpoints and Keypoints. In *Conference on Computer Vision and Pattern Recognition*, 2015. 2
- [43] Daniel Wagner, Gerhard Reitmayr, Alessandro Mulloni, Tom Drummond, and Dieter Schmalstieg. Pose Tracking from Natural Features on Mobile Phones. In *International Symposium on Mixed and Augmented Reality*, September 2008. 1
- [44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2, 6
- [45] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems*, 2018. 1, 2, 5, 6, 7, 8
- [46] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Åström, and Masatoshi Okutomi. Revisiting the PnP Problem: A Fast, General and Optimal Solution. In *International Conference on Computer Vision*, 2013. 2
- [47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. In *arXiv Preprint*, 2017. 6